BIOLOGICAL CRITERIA

Technical Guidance for Survey Design and Statistical Evaluation of Biosurvey Data

CHAPTER 2. Classical Statistical Interference and Uncertainty 3

CHAPTER 2 Classical Statistical Inference and Uncertainty

Before the biological survey can be designed and linked to statistical methods of interpretation, an exact formulation of the problem is needed to narrow the scope of the study and focus investigators on collecting the data. The choice of biological and chemical variables should be made early in the process, and the survey design built around that selection. Fancy statistics and survey designs may be appropriate, but biologically defined objectives should dominate and use the statistics, not the reverse (Green, 1979).

Formulating the Problem Statement

A clear statement of the objective or problem is the necessary basis on which the biological survey is designed. A general question such as "does the effluent from the municipal treatment plant damage the environment?" does little to help decision makers. Consider, however, their response to a more specific statement: "Is the mean abundance of young-of-the-year green sunfish caught in seines above the discharge point greater (with an error rate of 5 percent) than those similarly trapped downstream of the discharge point?" The precise nature of this question makes it a clear guide for the collection and interpretation of data.

The problem statement should minimally include the biological variables that indicate environmental damage, a reference to the comparisons used to determine the impact, and a reference to the level of precision (or uncertainty) that the investigator needs to be confident that an impact has been determined. In the preceding example, green sunfish are the biological indicator of impact, upstream and downstream seine data are the basis of comparison, and an error rate of 5 percent provides an acceptable level of uncertainty.

The problem statement, the survey design, and the statistical methods used to interpret the data are closely linked. Here, the survey design is an upstream/downstream set of samples with the upstream data providing a reference for comparison. A *t* test or rank sign test may be used to test for mean differences between the sites.

From a statistical standpoint, the biological variables (measures) used to show damage should have low natural variability and respond sharply to an impact relative to any sampling variability. Natural variability contributes to the uncertainty associated with their response to an impact. Lower natural variability permits reliable inferences with smaller sample sizes.

Examining historical data is an excellent means of selecting biological criteria that are sensitive to environmental impacts. Species that exhibit large natural spatial and temporal variations may be suitable indicators of environmental change only in small time scales or localized areas. If so, the use of such variables will limit the investigator's ability to assess environmental change in long-term monitoring programs. Historical data, combined with good scientific judgment, can be used to select biological criteria that exhibit minimal natural variability within the context of the site under evaluation.

Basic Statistics and Statistical Concepts

When a data set is quite small, the entire set can be reported. However, for larger data sets, the most effective learning takes place when investigators summarize the data in a few well-chosen statistics. The choice to trade some of the information available in the entire set for the convenience of a few descriptive statistics is usually a good one, provided that the descriptive statistics are carefully selected and correctly represent the original data.

Some descriptive statistics are so commonly used that we forget that they are but one option among many candidate statistics. For example, the mean and the standard deviation (or variance) are statistics used to estimate the center of a data set and the spread on those data. The scientist who uses these statistics has already decided that they are the best choices to describe the data. They work very well, for example, as representatives of symmetrically distributed data that follow an approximately normal distribution. Thus, their use in such circumstances is entirely justified. However, in other situations involving biological data, alternative descriptive statistics may be preferred.

Descriptive Statistics

Before selecting a descriptive statistic, the scientist must understand the purpose of the statistic. Descriptive statistics are often used in biological studies because the convenience of a few summary numbers outweighs the loss of information that results from not using the entire data set. Nevertheless, as much information as possible must be summarized in the descriptive statistics because the alternative may involve a misrepresentation of the original data.

The basic statistics and statistical techniques used in this chapter are further defined, described, and illustrated in the appendix to this document (Appendix A). Readers unfamiliar with descriptive statistics and graphic techniques should read Appendix A now and use it hereafter as a reference. Other readers may proceed directly to the tables in this chapter, which summarize the advantages and disadvantages of the statistical estimators and techniques described in the appendix.

The common measures of the center, or central tendency, of a data set are the mean, median, mode, geometric mean, and trimmed mean. None of these options is the best choice in all situations (see Table 2.1), yet each conveys useful information. The points raised in Table 2.1 are not comprehensive or absolute; they do, however, reflect the author's experience with these estimators.

Environmental contaminant concentration data are strictly positive, and sample data sets exhibit asymmetry (i.e., a few relatively high observations). Therefore, a transformation, in particular, the logarithmic transformation, should be applied to concentration and other data that exhibit these characteristics before analysis. When a transformation is used, data analysis and estimation occur within the transformed metric; if appropriate, the results may be converted back to the original metric for presentation.

A measure of dispersion — spread or variability — is another commonly reported descriptive statistic. Common estimators for dispersion are standard deviation, absolute deviation, interquartile range, and range. These estimators are defined, described, and illustrated with examples in the appendix; Table 2.2 summarizes when and how they may be used.

Table 2.3 summarizes four of the most useful univariate and bivariate graphic techniques, including histograms, stem and leaf displays, box and whisker plots, and bivariate plots. These methods are also illustrated in Appendix A.

Recommendations

There is no rigorous theoretical or empirical support for using the normal distribution as a population model for chemical and biological measures of water quality or as a model for errors. Instead, the evidence supports using the lognormal model. However, uncertainty about the correctness of the lognormal model suggests that prudent investigators will recommend estimators that perform well even if an assumed model is wrong.

Table 2.1—Measures of central tendency.

ESTIMATOR	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN	SHOULD NOT USE WHEN
Mean	• Most widely known and used choice • Easy to explain	• Not resistant to outliers • Not as efficient[1] as some alternatives under deviations from normality	• Sample mean is required • Distribution is known to be normal • Distribution is symmetric	• Outliers may occur • Distribution is not symmetric
Median	• Easy to explain • Easy to determine • Resistant to others	• Not as efficient as the mean under normality	• Sample median is required • Outliers may occur	
Mode	• Easy to explain • Easy to determine	• Not as efficient as the mean under normality	• Most frequently observed value is required • Data are discrete or can be discretized	• More efficient options are appropriate
Geometric Mean	• Appropriate for certain skewed (lognormal) distribution	• Not as easy to explain as first three	• Distribution appears lognormal	• More widely known estimators are appropriate
Trimmed Mean	• Resistant to outliers	• Not as easy to explain as first three	• Outliers may occur and estimator efficiency is desired	

[1] Higher efficiency means lower standard error.

Table 2.2—Measures of dispersion.

ESTIMATOR	ADVANTAGES	DISADVANTAGES	SHOULD CONSIDER FOR USE WHEN	SHOULD NOT USE WHEN
Standard Deviation	• Most widely known • Routinely calculated by statistics packages	• Strongly influenced by outliers • Not as efficient[1] as some alternatives under even slight deviations from normality	• Sample standard deviation is required • Distribution is known to be normal	• Outliers may occur • Sample histogram is even slightly more dispersed than is a normal distribution
Median Absolute Deviation	• Resistant to outliers	• Not as efficient as the standard deviation under normality	• Outliers may occur	
Interquartile Range	• Resistant to outliers • Relatively easy to determine	• Not as efficient as the standard deviation under normality	• Outliers may occur	
Range	• Easy to determine	• Not as efficient	• Range is required	• Any of the above options is appropriate

[1] Higher efficiency means lower standard error.

Table 2.3—Useful graphic techniques.

TECHNIQUE	FEATURES	USEFUL FOR
Histogram	• Bar chart for data on a single (univariate) variable • Shows shape of empirical distribution	• Visual identification of distribution shape, symmetry, center, dispersion, and outliers
Stem and Leaf Display	• Same as histogram • Presents numeric values in display	• Same as histogram
Box and Whisker Plot	• Display of order statistics (extremes, quartiles, and median) • May be used to graph the same characteristic (e.g., variable) for several samples (e.g., different sampling sites)	• Visual identification of distribution shape, symmetry, center, dispersion, and outliers (single sample) • Comparison of several samples for symmetry, center, and dispersion
Bivariate Plot	• Scatter plot of data points (variable *x* versus variable *y*)	• Visual assessement of the strength of a linear relationship between two variables • Evidence of patterns, nonlinearity and bivariate outliers

Many books and articles have been written recently concerning the theoretical and empirical evidence in favor of nonparametric methods and robust and resistant estimators. Books that consider alternative estimators of center and dispersion (e.g., Huber, 1981; Hampel et al. 1986; Rey, 1983; Barnett and Lewis, 1984; Miller, 1986; Staudte and Sheather, 1990) build a strong case for more robust estimators than the mean and variance. Indeed, there is good evidence (Tukey, 1960; Andrews et al. 1972) that the mean and variance may be the worst choices among the common estimators for error-contaminated data. Several articles that involve comparisons of estimators on real data (e.g., Stigler, 1977; Rocke et al. 1982; Hill and Dixon, 1982) also favor robust estimators over conventional alternatives.

As a consequence, the median and the trimmed mean are recommended for the routine calculation of a data set's central tendency. The interquartile range and the median absolute deviation are recommended for calculation of the dispersion. These suggestions represent a compromise between robustness, ease of explanation, and calculation simplicity. For the trimmed mean, recommended amounts of trimming range from 10 percent (Stigler, 1977) to over 20 percent (e.g., Rocke et al. 1982). A critical argument in support of the trimmed mean is that interval estimation and hypothesis testing are still possible using the *t* statistic (Tukey and McLaughlin, 1963; Dixon and Tukey, 1968; Gilbert, 1987).

Uncertainty

In statistics, uncertainty is a measure of confidence. That is, uncertainty provides a measure of precision — it assigns the value of scientific information in ecological studies. Scientific uncertainty is present in all studies concerning biological criteria, but uncertainty does not prevent management and decision making. Rather, uncertainty provides a basis for selecting among alternative actions and for deciding whether additional information is needed (and if so, what experimentation or observation should take place).

In ecological studies, scientific uncertainty results from inadequate scientific knowledge, natural variability, measurement error, and sampling error (e.g., the standard error of an estimator). In the actual analysis, uncertainty arises from erroneous specification of a model or from errors in statistics, parameters, initial conditions, inputs for the model, or expert judgment.

In some situations, uncertainty in an unknown quantity (e.g., a model parameter or a biological endpoint) may be estimated using a measure of variability. Likewise, in some situations, model error may be estimated using a measure of goodness-of-fit (predictions versus observations) of the model. In many situations, a judicious estimate of uncertainty is the only option; in these cases, careful estimation is an acceptable alternative and methods exist to elicit these judgments from experts (Morgan and Henrion, 1990).

In many studies, uncertainty is present in more than one component (e.g., parameters and models), so the investigator must estimate the combined effects of the uncertainties on the endpoint. This exercise, called error propagation, is usually undertaken with Monte Carlo simulation or first-order error analysis.

The outcome of an uncertainty analysis is a probability distribution that reflects uncertainty on the endpoint. However, uncertainty analysis may not always be the most useful expression of risk. Other expressions of uncertainty, such as prediction, confidence intervals, or odds ratios are easier to understand and interpret. If important error terms are ignored when a probability statement is made, the investigator must report this omission. Otherwise, the probability statement is not representative, and the uncertainties are underestimated.

Since uncertainty provides a measure of precision or value, it can be used by decision makers to guide management actions. For example, in some cases the uncertainty in a biological impact may be too large to justify management changes. As a consequence, managers may defer action until additional monitoring data can be gathered rather than require pollutant discharge controls. If the uncertainty is large and the estimated costs of additional pollutant controls quite high, it may be wise either to defer action or to look for smaller, relatively less expensive abatement strategies for an interim period while the monitoring program continues.

Though environmental planners at national, state, and local levels have rarely considered uncertainty in their planning efforts, their work has been generally successful over the past 20 years. It is, however, certainly possible that more effective management — that is, less costly, more beneficial management — might have occurred if uncertainty had been explicitly considered.

If overall uncertainty is ignored, the illusion prevails that scientific information is more precise than it actually is. As a consequence, we are surprised and disappointed when biological outcomes are substantially different from predictions. Moreover, if we don't calculate uncertainty, we have no rational basis for specifying the magnitude of our sampling program or the resources (money, time, personnel) that should be allocated to planning. Thus, decisions on planning and analysis are more likely based on convention and whim than on the logical objective of reducing scientific uncertainty.

Statistical analysis is largely concerned with uncertainty and variability. Therefore, uncertainty is an important concept in this guidance manual. The analyses presented here and in subsequent chapters are based on particular measures of uncertainty, for example, confidence intervals. These measures are "statistics"; they reflect data, and are not always considered in the broader context of uncertainty — that is, as establishing the uncertainty in a quantity of interest. We will, however, consider these statistics in the broader sense, with concern for the theoretical issues raised in this section. Particularly given the small samples that often occur with biocriteria assessments, investigators should ask the following questions:

- Do the data adequately represent uncertainty?
- Are all important sources of uncertainty represented in the data?
- Should expert scientific judgment be used to augment or correct measures of uncertainty?
- If components of uncertainty are ignored because they are not included in the data, are conclusions or decisions affected?

Statistical analysis is not a rote exercise devoid of judgment.

Statistical Inference

Statistical inference is gained by two primary approaches: (1) interval estimation, and (2) hypothesis testing. Interval estimation concerns the calculation of a confidence interval or prediction interval that bounds the range of likely values for a quantity of interest. The end product is typically the estimated quantity (e.g., a mean value) plus or minus the upper and lower interval. The same information is used in hypothesis testing; however, in hypothesis testing, the end product is a decision concerning the truth of a candidate hypothesis about the magnitude of the quantity of interest.

In a particular problem, the choice between using interval estimation or hypothesis testing generally depends on the question or issue at hand. For example, if a summary of scientific evidence is requested, confidence intervals are apt to be favored; however, if a choice or decision is to be made, hypothesis tests are likely to be preferred.

Interval Estimation

Statistical intervals, whether confidence or prediction, may be based on an assumed probability model describing the statistic of interest, or they may require no assumption of a particular underlying probability model.

Hahn and Meeker (1991) note that the proper choice of statistical interval depends on the problem or issue of concern. As a rule, if the interval is intended to bound a *population* parameter (e.g., the true mean), then the appropriate choice is the confidence interval. If, however, the interval is to bound a *future* member of the population (e.g., a forecasted value), then the appropriate choice is the prediction interval. Another statistical interval less frequently used in ecology is the tolerance interval, which bounds a specified *proportion* of observations.

In conventional (classical, or frequentist) statistical inference, the statistical interval has a particular interpretation that is often incorrectly stated in scientific studies. For example, if a 95 percent statistical interval for the mean is 7 ± 2, it is not correct to say that there is a 95 percent chance that the true mean lies between 5 and 9." Rather, it is correct to say that 95 percent of the time this interval is calculated, the true mean will lie within the computed interval. Although it sounds awkward and not directly relevant to the issue at hand, this interpretation is the correct meaning of a classical statistical interval. In truth, once it is calculated, the interval either does or does not contain the true value. In classical statistics, the inference from interval estimation refers to the procedure for interval calculation, not to the particular interval that is calculated.

Hypothesis Testing

Biosurveys are used for many purposes, one of which is to assess impact or effect. Resource managers may want to assess, for example, the influence of a pollutant discharge or land use change on a particular area. The effect of the impact can be determined based on the study of trends over time or by comparing upstream and downstream conditions. In some instances, the interest is in magnitude of effect, but concern often focuses simply on the presence or absence of an effect of a specific magnitude. In such cases, hypothesis testing is usually the statistical procedure of choice.

In conventional statistical analysis, hypothesis testing for a trend or effect is often based on a point null hypothesis. Typically, the point null hypothesis is that no trend or effect exists. The position is presented as a "straw man" (Wonnacott and Wonnacott, 1977) that the scientist expects to reject on the basis of evidence. To test this hypothesis, the investigator collects data to provide a sample estimate of the effect (e.g., change in biotic integrity at a single site over time). The data are used to provide a sample estimate of a test statistic, and a table for the test statistic is consulted to estimate how unusual the observed value of the test statistic is if the null hypothesis is true. If the observed value of the test statistic is unusual, the null hypothesis is rejected.

Table 2.4—Possible outcomes from hypothesis testing.

STATE OF THE WORLD	DECISION	
	ACCEPT H_0	REJECT H_0
H_0 is True	Correct decision. Probability = 1 – α; corresponds to the *confidence level.*	Type I error. Probability = α; also called the significance level.
H_0 is Fale (H_1 is True)	Type II error. Probability = β	Correct decision. Probablity = 1 – β; also called power.

In a typical application of parametric hypothesis testing, a hypothesis, H_0, called the null hypothesis, is proposed and then evaluated using a standard statistical procedure like the *t* test. Competing with this null hypothesis for acceptance is the alternative hypothesis, H_1. Under this simple scheme, there are four possible outcomes of the testing procedure: the hypothesis is either true or false, and the test results can be accepted or rejected for each hypothesis (see Table 2.4).

The point null hypothesis is a precise hypothesis that may be symbolically expressed:

$$H_0: \theta_1 - \theta_2 = 0$$
$$H_1: \theta_1 - \theta_2 \neq 0$$

where θ is a parameter of interest. An example of a point null hypothesis in words is, "no change occurs in mean IBI after the new wastewater treatment plant goes on line." Symbolically, it is expressed as

$$H_0: \mu_1\text{-}\mu_2 = 0$$
$$H_1: \mu_1\text{-}\mu_2 \neq 0$$

where μ_1 is the "before" true mean and μ_2 is the "after" true mean. The test of the null hypothesis proceeds with the calculation of the sample means, $\bar{x}_1$ and $\bar{x}_2$. In most cases, the sample means will differ as a consequence of natural variability or measurement error or both, so a decision must be made concerning how large this difference must be before it is considered too large to result from variability or error. In classical statistics, this decision is often based on standard practice (e.g., a Type I error of 0.05 is acceptable), or on informal consideration of the consequences of an incorrect conclusion.

The result of a hypothesis test can be a conclusion or a decision concerning the rejected hypothesis. Alternatively, the result can be expressed as a "p-value," which quantifies the strength of the data evidence in favor of the null hypothesis. The p-value is defined as the probability that "the sample value would be as large as the value actually observed, if H_0 is true" (Wonnacott and Wonnacott, 1977). In effect, the p-value provides a measure of how likely a particular value is, assuming that the null hypothesis is true. Thus, the smaller the p-value, the less likely that the sample supports H_0. This is useful information; it suggests that p-values should always be reported to allow the reader to decide the strength of the evidence.

Common Assumptions

Virtually all statistical procedures and tests require the validity of one or more assumptions. These assumptions concern either the underlying population being sampled or the distribution for a test statistic. Since the failure of an assumption can have a substantial effect on a statistical test, the common assumptions of normality, equality of variances, and independence are discussed in this section. We must ask, for example, to what extent can an assumption be violated without serious consequences? Or how should assumption violations be addressed?

■ **Normality.** A common assumption of many parametric statistical tests is that samples are drawn from a normal distribution. Alternatively, it may be assumed that the statistic of interest (e.g., a mean) is described by a normal sampling distribution. In either case, the key distinction between parametric and nonparametric (or distribution-free) statistical tests is that a probability model (often normal) is assumed.

Empirical evidence (e.g., Box et al. 1978) indicates that the significance level but not the power is robust or not greatly affected by mild violations of the normality assumption for statistical tests concerned with the mean. This finding suggests that a test result indicating "statistical significance" is reliable, but a "nonsignificant" result may be the result of a lack of robustness to nonnormality. The normality of a sample can be checked using a normal probability plot, chi square test, Kolmogorov-Smirnov test, or by testing for skewness or kurtosis; however, many biological surveys are not designed to produce enough samples to make these tests definitive.

Normality of the sampling distribution for a test statistic is important because it provides a probability model for interval estimation and hypothesis tests. In some cases, transformation of the data may help the investigator achieve approximate normality (or symmetry) in a sample, if normality is required. Since nonnegative concentration data cannot be truly normal, and since empirical evidence suggests that environmental contaminant data may be described with a lognormal distribution, the logarithmic transformation is a good first choice. Therefore, in the absence of contrary evidence, we recommend that concentration data be log-transformed prior to analysis.

■ **Equality of Variance.** A second common assumption is that when two or more distributions are involved in a test, the variances will be constant across distributions. Many tests are also robust to mild violations of this assumption, particularly if the sample sizes are nearly identical. To test this assumption, a *t* test (usually a two-tailed one) can be performed; see Snedecor and Cochran (1967) for an example, and Miller (1986) for interpretive results. Conover (1980) provides an alternative, namely, nonparametric tests of equality of variances. Note that if two means are being compared based on samples with vastly different variances, the differences of interest may be more fundamental than the difference between the means.

■ **Independence.** The assumption of greatest general concern is independence. Most statistical tests (parametric and nonparametric) require a random sample, or a sample composed of independent observations. Dependency between or among observations in a data set means that each observation contains some information already conveyed in other observations. Thus, there is less new independent information in a dependent data set than in an independent data set of the same sample size. Because statistical procedures are often not robust to violation of the independence assumption, adjustments are generally recommended to address anticipated problems.

Dependence in a sample can result from spatial or temporal patterns, that is, from persistence through time and space. In most types of analyses, the assumption of independence refers to independence in the disturbances (errors). For example, in a time series with temporal trend and seasonal pattern, dependence or autocorrelation in the raw data series may exist because of a deterministic feature of the data (e.g., the time trend or seasonal pattern).

This type of autocorrelation poses no difficulty; it is addressed by modeling the deterministic features of the data and subtracting the modeled component from the original series. Of particular concern in testing for trend is autocorrelation that remains after all deterministic features are removed (i.e., errors that are in the disturbances). When this situation arises, an adjustment to the trend test is necessary. Reckhow et al. (1993) provide guidance and software.

A similar situation can occur in the estimation of a regression slope or a central tendency statistic such as the mean or trimmed mean. In such cases, the independence assumption refers to the errors, as estimated by the residuals, around the regression line or the mean. If persistence or dependence is found in the residuals, then the independence assumption is violated and corrective action is needed. Options to address this problem include using an effective sample size (Reckhow and Chapra, 1983), or generalized, least squares for regression (see Kmenta [1986] or any standard econometrics regression text).

If the investigator finds positive autocorrelation in the disturbances (i.e., if each disturbance is positively correlated with nearby disturbances in the series), confidence interval estimates will be too narrow and may lead to rejection of the null hypothesis. Autocorrelation in the disturbances is the most common and potentially the most troublesome of the causes of assumption violations.

The degree of autocorrelation is a function of the frequency of sampling; that is, a data set based on an irregular sampling frequency cannot be characterized by a single, fixed value for autocorrelation. For biological time series, stream data obtained more frequently than monthly may be expected to be autocorrelated (after trends and seasonal cycles are removed). Stream survey data based on less frequent sampling are less likely to exhibit sample autocorrelation estimates of significance.

Parametric Methods — the t Test

Parametric approaches involve a model (e.g., regression slope) for any deterministic features and a probability model for the errors. In some cases, the deterministic model will be a linear, curvilinear, or step function, while the model for the errors is typically a normal probability distribution with independent, identically distributed errors. In other cases, the deterministic model may simply be a constant (as it is when interest focuses on an "upstream/downstream" comparison between two sites), though the probability model may in all cases be a normal probability distribution. The *t* test is a typical parametric test.

Using the *t* test

A Student's *t* statistic:

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}} \qquad \textbf{(2.1a)}$$

has a Student's *t* distribution (n-1 degrees of freedom); here, "x" is the mean of a random sample from a normal distribution with true mean μ and constant variance, s is the sample standard deviation, and n is the sample size. In addition, for two samples:

$$t=\frac{\bar{x}_1-\bar{x}_2}{\frac{s_1+s_2}{2}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad \textbf{(2.1b)}$$

also has a Student's *t* distribution (n_1+n_2-2 degrees of freedom); here, x_1 and x_2 are the sample means; s_1 and s_2 are the sample standard deviations; and n_1 and n_2 are the sample sizes. This distribution is widely tabulated, and it is commonly used in hypothesis testing and confidence interval estimation for a sample mean (one-sample test; Equation 2.1a) or a comparison of sample means (two-sample test; Equation 2.1b).

When Student's *t* distribution is used in a hypothesis test (a *t* test), it is assumed that samples are drawn from a normal distribution, the variances are constant across distributions, and the observations are independent. Of these assumptions, Box et al. (1978) have shown that the *t* test has limited robustness to violations of the first two (normality and equality of variances); however, problems will occur if the observations are dependent. The scientist should probably be concerned about the first two assumptions only in situations in which the two data sets have substantially different variances and substantially different sample sizes (see Snedecor and

Cochran [1967] for *F* test calculations to compare variances).

An attractive variation of the *t* statistic for use in situations where outliers are of concern was proposed by Yuen and Dixon (1973; see also Miller, 1986; and Staudte and Sheather, 1990). They created an outlier-resistant, or robust, version of the *t* statistic (Equations 2.1a and 2.1b) using a trimmed mean and a Winsorized standard deviation. For example, if a *t* statistic is used to compare the means of two populations, the robust (trimmed *t*) version is

$$t_{tri} = \frac{\bar{x}_{tri1} - \bar{x}_{tri2}}{s_w \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.2}$$

where $\bar{x}_{tri}$ = trimmed mean for sample i

s_w = Winsorized standard deviation

n_i = number of observations in sample i

A Winsorized statistic is similar to a trimmed statistic. For trimming, observations are ordered from lowest to highest, and the *k*-lowest and *k*-highest are removed from the sample for the calculation of the *k*-trimmed statistic (e.g., trimmed mean). For *k*-Winsorizing, observations are ordered from lowest to highest, and the *k*-lowest and *k*-highest are not removed, but are reassigned the values of the lowest observation and the highest observation remaining in the trimmed sample. The following example illustrates this.

A sample of 10 IBI values is obtained for analysis:
9, 31, 26, 25, 34, 38, 33, 31, 28, 37

And ordered from lowest to highest:
25, 26, 28, 29, 31, 31, 33, 34, 37, 38.

The 10 percent-trimmed sample is
26, 28, 29, 31, 31, 33, 34, 37

The 10 percent-Winsorized sample is
26, 26, 28, 29, 31, 31, 33, 34, 37, 37.

If we were to calculate the 10 percent-trimmed *t* statistic in Equation 2.2 for this IBI sample, we would use: (1) the trimmed sample (eight observations) to calculate a mean, and (2) the Winsorized sample (10 observations) to calculate a standard deviation. For the two-sample comparison of means, the trimmed *t* statistic has $(1\text{-}2k)(n_1+n_2)-2$ degrees of freedom or, in the above example, 7 degrees of freedom (df). The trimmed *t* statistic is an attractive option that should be considered whenever outliers are a concern.

The parametric approach is appropriate and advantageous if the deterministic model is a reasonable characterization of reality and if the model for errors holds. In such cases, parametric tests should be more powerful than nonparametric or distribution-free alternatives. Thus, the assumption that deterministic and probability models are correct is the basis on which the superior performance of parametric methods rests. If the assumptions concerning these models are incorrect, then the results of the parametric tests may be invalid and distribution-free procedures may be more appropriate.

Nonparametric Tests — the W test

Distribution-free methods, as the name suggests, do not require an assumption concerning the particular form of the underlying probability model for the data generation process. An assumption of independence is, however, usually made; therefore, autocorrelation can be as serious a problem in nonparametric methods as it is for parametric and robust methods. Distribution-free tests are often based on rank (or order); the sample observations are arranged from lowest to highest. The Wilcoxon-Mann-Whitney test or *W* test is a typical distribution-free test.

Using the *W* test

The *W* test is a two-sample hypothesis test, designed to test the hypothesis that two random samples are drawn from identical continuous distributions with the same center (alternative hypothesis: one distribution is offset from the other, but otherwise identical). This test is often presented as an option to the two-sample *t* test that should be considered if the assumption of normality is believed to be seriously in error. The *W* test has its own statistic, which is tabulated in most elementary statistics textbooks (i.e., those with a chapter on nonparametric methods). However, for moderate to large sample sizes (e.g., $n > 15$), the statistic is approximately normal under the null hypothesis, so the standard normal table can be used.

The scientist should consider the *W* test for any situation in which the two-sample *t* test may be used. Comparative studies of these two tests indicate that while the *t* test is robust to violations of the normality assumption, the *W* test is relatively powerful while not requiring normality. Situations that appear severely nonnormal might favor the *W* test; otherwise the *t* test may be selected. Some statisticians (e.g., Blalock, 1972) recommend that both tests be conducted as a double check on the hypothesis.

Unfortunately, violation of the independence assumption appears to be as serious for the *W* test as for the *t* test. If these tests are to be meaningful, the scientist must confirm independence or make other adjustments as noted in Reckhow et al. (1993).

In essence, the *W* test is used to determine if the two distributions under study have the same central tendency, or if one distribution is offset from the other. To conduct the *W* test, the data points from the samples are combined, while maintaining the separate sample identity. This overall data set is ordered from low value to high value, and ranks are assigned according to this ordering.

To test the null hypothesis of no difference between the two distributions ($f\,[x]$ and $g[x]$)

$$H_0\colon f(x) = g(x)$$

the ranks, R_i, for the data points in one of the two samples are summed:

$$W = \sum R_i \qquad (2.3)$$

The ranks should be specified as follows (Wonnacott and Wonnacott, 1977): Start ordering (low to high, or high to low) from the end (high or low) at which the observations from the smaller sample tend to be greater in number, and sum the ranks to estimate *W* from this smaller sample. This estimate keeps *W* small as it is reported in most tables. For either one-sided or two-sided tests, if ties occur in the ranks, then all tied observations should be assigned the same average rank.

Statistical significance is a function of the degree to which, under the null hypothesis, the ranks occupied by either data set differ from the ranks expected as a result of random variation. For small samples, the *W* statistic calculated in Equation 2.3 can be compared to tabulated values to determine its significance (see Hollander and Wolfe, 1973). For moderate to large samples (where total *n* from both samples > 15), *W* is approximately normal (if the null hypothesis is true). Therefore, the *W* statistic may be evaluated using a standard normal table with mean ($E[W]$) and variance ($\mathrm{Var}[W]$):

$$E(W) = n_A\,(n_B + n_A + 1)/2 \qquad (2.4)$$

$$\mathrm{Var}(W) = n_B n_A (n_B + n_A + 1)/12 \qquad (2.5a)$$

If there are ties in the data, then the variance may be calculated as

$$Var(\mathrm{W}) = \frac{\mathrm{n_A n_B}}{12}\left[\mathrm{n_A} + \mathrm{n_B} + 1 - \frac{\sum_{j=1}^{g} t_1(\mathrm{t}_j^2 - 1)}{(\mathrm{n_A} + \mathrm{n_B})(\mathrm{n_A} + \mathrm{n_B} - 1)}\right] \qquad (2.5b)$$

where t_j is the size (number of data points with the same value) of tied group *j*. The effect of ties is negligible unless there are several large groups ($t_j \geq 3$) in the data set.

These statistics are used to create the standard normal deviate:

$$z = \frac{\mathrm{W} - \mathrm{E(W)}}{(Var(\mathrm{W}))^{0.5}} \qquad (2.6)$$

where: n_A, n_B = the number of observations in samples *A* and *B* ($n_A < n_B$).

Example — an IBI case study

IBI data have been obtained from upstream and downstream sites surrounding a wastewater discharge. Assume independence.

Upstream	33	34	2.5	3.7	39	45	49	47	45	44
Downstream	26	30	18	32	36	36	43	42	41	41

(a) Test the null hypothesis that the true difference between the upstream and downstream IBI means is zero, versus the alternative hypothesis that the downstream IBI mean is *lower* than the upstream IBI mean.

$$H_0\colon \mu_U - \mu_D = 0$$

$$H_1\colon \mu_U - \mu_D > 0$$

First, some basic statistics for each sample:

	SAMPLE MEAN	SAMPLE STANDARD DEVIATION
Upstream	39.8	7.57
Downstream	34.5	8.09

For a comparison of two means based on equal sample sizes, the *t* statistic is

$$\mathrm{t} = \frac{\bar{\mathrm{x}}_1 - \bar{\mathrm{x}}_2}{\frac{(\mathrm{s}_1 + \mathrm{s}_2)}{2}\sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}}} = \frac{39.8 - 34.5}{\frac{7.57 + 8.09}{2}\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{53}{7.83\sqrt{0.2}} = \frac{53}{35} = 1.51$$

At the 0.05 significance level, the one-tailed *t* statistic for 18 degrees of freedom is 1.73. Since 1.51 < 1.73, we cannot reject the null hypothesis (at the 0.05 level).

(b) Test the null hypothesis (see part a) using the 10 percent trimmed *t* (10 percent trimmed from each end).

$$\mathrm{t_{tri}} = \frac{\bar{\mathrm{x}}_{\mathrm{tri1}} - \bar{\mathrm{x}}_{\mathrm{tri2}}}{\frac{({}^{s}\mathrm{w1} + {}^{s}\mathrm{w2})}{2}\sqrt{\frac{1}{\mathrm{n}_1} + \frac{1}{\mathrm{n}_2}}} = \frac{40.5 - 35.5}{\frac{5.83 + 6.39}{2}\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{5.0}{6.12\sqrt{0.2}} = 1.83$$

At the 0.05 significance level, the one-tailed *t* statistic for 14 degrees of freedom is 1.76. Since 1.83 1.76, we reject the null hypothesis (at the 0.05 level).

(c) Test the null hypothesis (see part a) using the *W* test.

ORDER IBI VALUES																				
Upstream	49	47	45	45	44					39	37			34	33				25	
Downstream						43	42	41	41			36	36			32	30	26		18
ORDER																				
Upstream	1	2	3.5	3.5	5					10	11			14	15				19	
Downstream						6	7	8.5	8.5			12.5	12.5			16	17	18		20

Here the separate samples have been combined for the purpose of rank ordering. The *W* test statistic can then be calculated from the ranks:

$$W = \sum R_{Up} = 1+2+3.5+3+5+10+11+14+15+19 = 84$$

$$E(W) = n_A(n_B + n_A + 1)/2 = 10(10+10+1)/2 = 105$$

$$Var(W) = (n_B n_A / 12)[n_B + n_A + 1 - \{\sum t^2 - 1)\} / (n_B + n_A)(n_B + n_A - 1)]$$

$$= [(10)(10)/12][10 + 10 + 1 - \{(2)(3) + (2)(3) + (2)(3)\} / (10 + 10)(10 + 10 - 1)] = 174.61$$

$$z = \frac{(W - E(W)}{(Var(W))^{0.5}} = \frac{84 - 105}{174.61^{0.5}} = -1.59$$

At the 0.05 significance level, the one-tailed *z* statistic is 1.65. Since 1.59 < 1.65, we cannot reject the null hypothesis (at the 0.05 level.

A glance at the IBI values and ranks in this example indicates a difference between the two samples (box plots and histograms would provide further supporting evidence). At issue is whether this difference in the sample is a chance occurrence or an indication of a true difference between the sites. If we adopt the conventional 0.05 level for hypothesis testing, then the conclusions from the three tests are ambiguous. Still, we can say the following about both the site comparisons and the methods:

(i) The downstream site is slightly impacted. Even though only one of the three test results yielded significance (at the 0.05 level), all three were close, suggesting a slight difference between the sites.

(ii) For each site, the lowest IBI value (25 for upstream, and 18 for downstream) is influential, particularly on the standard deviation. As a consequence, for the conventional *t* test, the denominator in the *t* statistic is inflated and rejection of the null hypothesis is less likely. Note that the lowest IBI value for the upstream site (IBI = 25) also affects the distribution-free *W* test. This IBI value holds a high rank (19) for the upstream sample, and substantially affects the test result. If that single IBI value had been 27 instead of 25, we would have rejected the null hypothesis at the 0.05 level.

(iii) The trimmed *t* is resistant to unusual observations or outliers, and thus provides the best single indicator of difference between the sites as conveyed by the bulk of the data from each site.

Conclusions

In hypothesis testing, the conclusion to not reject H_0 (in effect, to accept H_0) should not be evaluated strictly on the basis of *α*, the probability of *rejecting* H_0 when it is true (Type I error; see Table 2.4). Instead, we must be concerned with β, the probability of accepting H_0 when it is false (Type II error). Unfortunately, β does not have a single value, but is dependent on the true (but unknown) value of the difference between population means and on the sample size, *n*. For a particular testing procedure and sample size, we can determine and plot a relationship between the true difference between means and β. This plot is called the operating characteristic curve.

To understand the issues concerning significance and power (α and 1-β), consider the null hypothesis in the IBI case study:

H_0: The population mean IBI at the upstream site is the same as the population mean IBI at the downstream site.

In addition, because of the wastewater discharge, consider the general alternative hypothesis:

H_A: The population mean IBI at the upstream site is higher than the population mean IBI at the downstream site.

If we adopt α = 0.05 (the probability of rejecting H_0 when it is true; Type I error) as our significance level, then Figure 2.1a displays the sampling distribution for the mean under H_0 with 18 degrees of freedom. The horizontal axis in Figure 2.1 is the "difference between the means"; thus, the sampling distribution is centered at zero in Figure 2.1a (consistent with zero difference between means under H_0). The 0.05-significant tail area (the "rejection region") begins at 6.06, which means that the sample difference must be greater than or equal to 6.06 for us to reject H_0. Since the difference between the means in our sample IBI was only 5.3, we are inclined to accept the null hypothesis, based on the conventional *t* test.

(Note: to find the beginning of the tail area multiply the *t* statistic times the standard error. In this example, the *t* statistic is 1.73 [one-sided, 0.05 level, 18 degrees of freedom], and the standard error is 3.5. Thus, the tail area begins at [1.73][3.5] = 6.06.)

Now suppose that the following alternative hypothesis, H_1, is actually true for the sample IBI case:

H_1: The population mean IBI at the upstream site is higher by 5.0 than the population mean IBI at the downstream site.

In addition suppose that while H_1 actually is true, we propose a hypothesis test for H_0 based on the acceptance region in Figure 2.1a (i.e., accept H_0 if the

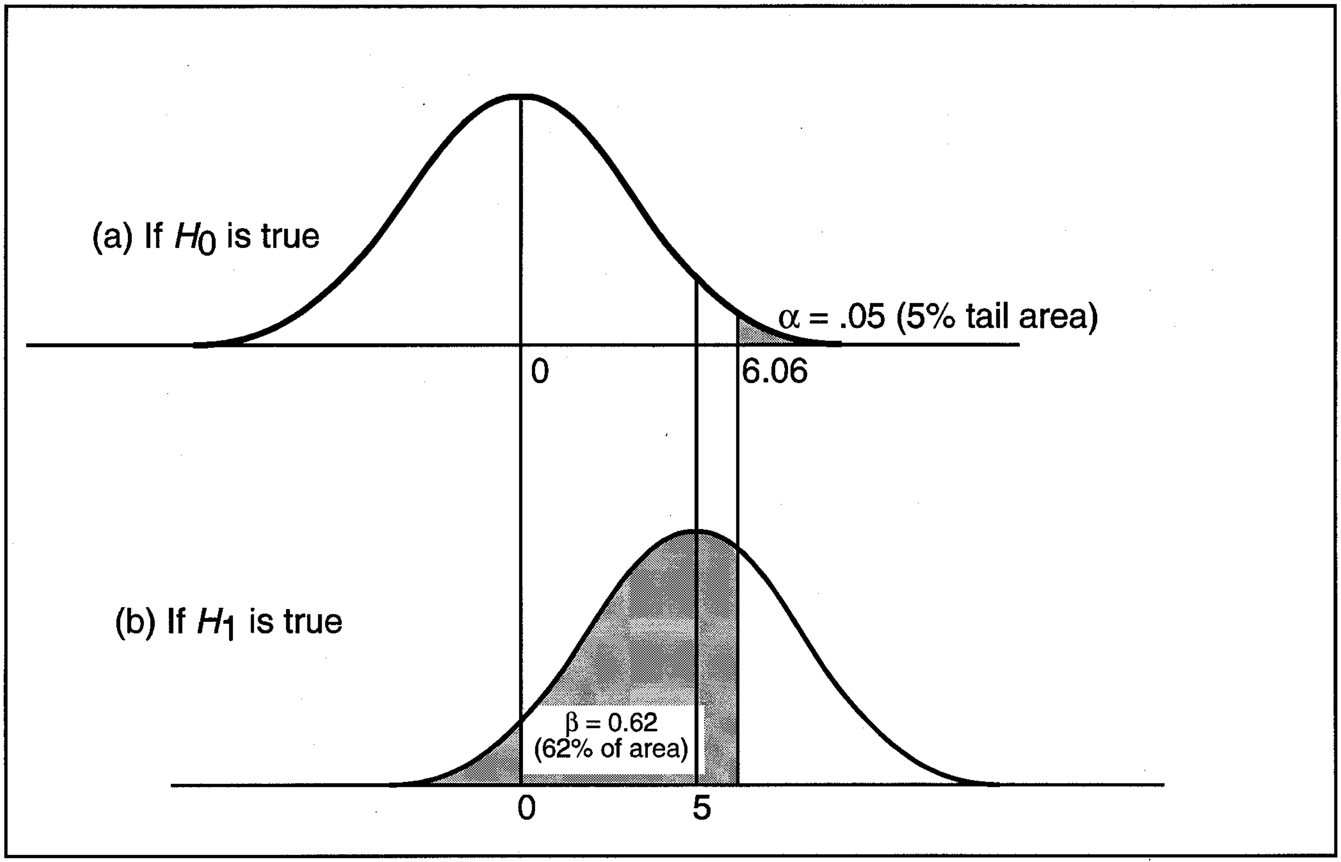

Figure 2.1a and b.—Sampling distributions under different hypotheses.

difference between the means is less than 6.06), which is exactly what occurred in our example. As we noted above, consideration of H_0 alone (Figure 2.1a) leads us to accept the null hypothesis.

Yet, with H_1 actually true (see Fig. 2.1b), if we propose a hypothesis test for H_0 based on the acceptance region in Figure 2.1a, there is a 62 percent chance that we will accept H_0 when it is actually false, according to Figure 2.1b (given the sample size in the example). This high likelihood of Type II error (see Table 2.4) underscores the danger of concluding the hypothesis test with acceptance of the null hypothesis. The power of this particular test is 1-β, or a 38 percent chance of detecting an IBI change of 5. Note that the specific alternative hypothesis H_1 is one example of an unlimited number of possibilities associated with the general alternative hypothesis H_A. Associated with H_1; $\beta = 0.62$ is one point on the power curve for this test and sample size. To properly determine the power of a test, we need to calculate β for a range of specific alternative hypotheses.

A second issue of concern in hypothesis testing is the problem of multiple simultaneous hypothesis testing, or "multiplicity" (Mosteller and Tukey, 1977). The classical interpretation of the 0.05 significance level (for α) associated with a hypothesis test is that 95 percent of the time this testing procedure is applied, the conclusion to accept the null hypothesis will not be in error if the null hypothesis is true. That is, on the average, one in 20 tests under these conditions will result in Type I errors.

The problem of multiplicity arises when an investigator conducts several tests of a similar nature on a set of data. If all but a few of the tests yield statistically insignificant results, the scientist should not ignore this in favor of those that are significant. The error of multiplicity results when one ignores the majority of the test results and cites only those that are apparently statistically significant. As Mosteller and Tukey (1977) note, the multiplicity error is technically the incorrect assignment of an α-level. When multiple tests of a similar nature are run on a set of data, a collective α should be used, associated with simultaneous test results. This tactic is typically referred to as the Bonferroni correction for correlation analysis.

The following comments from Wonnacott and Wonnacott (1972, pp. 201-202) summarize our attitude toward hypothesis testing:

> We conclude that although statistical theory provides a rationale for rejecting H_0, it pro-

vides no formal rationale for accepting H_0. The null hypothesis may sometimes be uninteresting, and one that we neither believe or wish to establish; it is selected because of its simplicity. In such cases, it is the alternative H_1 that we are trying to establish, and we prove H_1 by rejecting H_0. We can see now why statistics is sometimes called "the science of disproof." H_0 cannot be proved, and H_1 is proved by disproving (rejecting) H_0. It follows that if we wish to prove some proposition, we will often call it H_1 and set up the contrary hypothesis H_0 as the "straw man" we hope to destroy. And of course if H_0 is only such a straw man, then it becomes absurd to accept it in the face of a small sample result that really supports H_1.

Since there are great dangers in accepting H_0, the decision instead should often be simply to "not reject H_0," i.e., reserve judgment. This means that type II error in its worse form may be avoided; but it also means you may be leaving the scene of the evidence with nothing in hand. It is for this reason that either the construction of a confidence interval or the calculation of a prob-value is preferred, since either provides a summary of the information provided by the sample, useful to sharpen up your knowledge of what the underlying population is really like.

If, on the other hand, a simple accept-or-reject hypothesis test is desired, then we must look to a far more sophisticated technique. Specifically, we must explicitly take account not only of the sample data used in any standard hypothesis test (along with the adequacy of the sample size), but also:

1. *Prior belief.* How much confidence do we have in the engineering department that has assured us that the new process is better? Is their vote divided? Have they ever been wrong before?

2. *Loss involved in making a wrong decision.* If we make a type I error (i.e., decide to reject the old process in favor of the new, even though the old is as good), what will be the costs of retooling, etc.?

These comments amount to an advocacy of Bayesian decision theory. While it may be difficult to interpret a biosurvey in decision analysis terms, prior information and loss functions should, at a minimum, be considered in an informal manner. It is good engineering and planning practice to make use of all relevant information in inference and decision making.